LA CONSTITUTION DE CORPUS EN DIACHRONIE LONGUE : METHODOLOGIES, OBJECTIFS ET EXPLOITATIONS LINGUISTIQUES ET STYLISTIQUES

English version below

Julie SORBA, Univ. Grenoble Alpes, LiDiLEM

Olivier KRAIF, Univ. Grenoble Alpes, LiDiLEM

Adam RENWICK, Univ. Grenoble Alpes, LiDiLEM

Corinne DENOYELLE, Univ. Grenoble Alpes, Litt&Arts UMR 5316

Depuis plusieurs décennies, la numérisation des textes anciens et les progrès du TAL pour les traiter et les interroger ont largement modifié nos habitudes de travail. Il est désormais possible d'obtenir des données quantitatives massives qui affinent notre perception des phénomènes linguistiques ou stylistiques pour des corpus écrits dans des états de langue anciens. Les corpus numériques créés depuis maintenant près d'un quart de siècle permettent d'envisager plus facilement la dynamique du français en diachronie longue dont l'aboutissement, après de nombreuses années de travail, de la *Grande Grammaire Historique du Français* (Marchello-Nizia *et alii*, 2020) constitue un bel exemple. Nous définissons un corpus en diachronie longue comme un corpus périodisé, regroupant des textes choisis pour leur caractère représentatif des états de langue (de l'ancien français au français contemporain) des périodes couvertes par le corpus.

Depuis les années 1980, les chercheurs et chercheuses bénéficient de la base textuelle *Frantext*, la première en langue française, qui a permis de mener des investigations, au sein de textes littéraires, sur un très large empan temporel. Le travail pionnier de la *Base de Français Médiéval* (1989, BFM) a permis la constitution d'un corpus de textes littéraires et non littéraires, toutefois limité, comme son nom l'indique, à la période de l'ancien français et du moyen français. La livraison d'une nouvelle version de la BFM est d'ailleurs prévue fin 2022. De très nombreux corpus plus spécifiques à un genre textuel ou à un état de langue les ont rejoints : par exemple, le corpus sur 6 siècles de coutumiers normands du projet *Condé*, le corpus de référence du français médiéval (SRCMF) ou le corpus de sermons protestants du 16e au 18e siècle du projet *Sermo*). Cette dynamique actuelle dans la constitution et l'exploitation des corpus diachronique était, par exemple, au cœur du colloque international *ConCorDiaL Constitution de corpus en diachronie longue*, organisé à l'Université Grenoble Alpes (octobre 2022), et sera abordée, en avril 2023 dans la journée d'étude *Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques* (Université Grenoble Alpes / Université Roma La Sapienza) et en 2024 dans la conférence *Tracing the Curve of evolution : Syntactic change through text types* (Université de Caen, mars 2024).

La première étape dans la construction d'un corpus, comme le rappellent Reppen (2010 : 31) et Nelson (2010 : 53), est de savoir précisément quel est l'objectif poursuivi. Par exemple,

la sélection de sources comparables pour permettre des analyses quantitatives homogènes est essentielle et la temporalité prise en compte dépend des phénomènes que l'on veut observer (*GGHF* 2020 : 43). Ensuite, la construction d'un corpus est le fruit de choix raisonnés qui visent à satisfaire le principe de la représentativité : « [a corpus is] a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. » (Tognini-Bonelli, 2001 : 2). Ce principe de représentativité recouvre des réalités diverses en fonction des objectifs visés par celles et ceux qui construisent les corpus : les lexicographes qui souhaitent rendre compte du sens d'unités lexicales n'auront pas les mêmes exigences de représentativité que les linguistes et stylisticiens qui travaillent sur la caractérisation d'un genre textuel. Certains posent comme essentiels le fait de recourir exclusivement à des textes intégraux (Rastier, 2011 : 33), d'autres rappellent qu'un corpus ne peut être qu'un échantillon et qu'à ce titre, il peut être construit à partir d'échantillons (Renouf, 1987 ; Biber, 1993). Ainsi, l'objectif de ce numéro de la revue *Corpus* est d'interroger d'une part les choix présidant à la création des corpus en diachronie longue, d'autre part les objectifs linguistiques mais aussi stylistiques ou littéraires qui déterminent leur constitution.

Les axes thématiques que nous proposons peuvent se situer dans une perspective à la fois rétrospective (quel a été l'apport des corpus diachroniques ? comment valoriser les corpus constitués au cours des dernières décennies ?) et prospective (quels sont les défis théoriques et méthodologiques qui attendent la recherche en diachronie à l'ère des humanités numériques et des corpus outillés ?). Les réflexions pourront s'appuyer sur des corpus en langue française ou en langue étrangère.

Axe 1 : La constitution d'un corpus

Créer des corpus aptes à fournir des données en diachronie longue pose de nouvelles questions d'homogénéité des outils et des supports à tous les niveaux de la chaine de préparation du matériau : de la sélection des textes à leur traitement. Par exemple, dans la présentation des critères choisis pour construire le corpus de la *GGHF* (2020 : 42-43), Prévost oppose d'un côté les textes sélectionnés selon des critères *paratextuels*, « qui relèvent davantage du point de vue que le locuteur moderne porte sur ces textes » et qui impliquent de choisir des textes de référence comme la *Chanson de Roland* ou la *Queste del Saint Graal*, et d'un autre côté les critères *descripteurs* qui relèvent plutôt de la temporalité propre à chaque phénomène linguistique. On interrogera en particulier

- la diversité ou l'homogénéité des textes, selon différents niveaux hiérarchiques (domaines, discours, genres ; sur ces catégories, voir par exemple, Malrieu & Rastier, 2001 ; Marchello-Nizia et *alii*, 2020) ou différentes variétés du français (diatopiques ou diastratiques) ;
- l'origine des textes que l'on veut y inclure selon que l'on s'appuie sur des sources secondaires (textes déjà édités) ou primaires (des textes restant à éditer). Si l'on privilégie des textes déjà édités, comment compenser l'inévitable hétérogénéité des choix éditoriaux ? Pour les sources primaires, quels choix éditoriaux effectuer sur le plan graphique (sachant que les traditions philologiques d'édition de textes diffèrent selon les siècles considérés en ce qui concerne par exemple la segmentation des mots,

la graphie, les accents, la ponctuation, les majuscules) ?

- les types de codage mis en place dans le traitement des textes (quelles informations additionnelles ont été privilégiées pour l'enrichissement des textes ? combien de couches d'annotations ont été choisies ?)

Axe 2 : Effectuer des recherches avec les corpus constitués

L'objectif d'un corpus influe sur sa constitution, il importe alors de s'interroger sur les données qu'on souhaite en extraire.

- Quel type de recherches permettent les corpus en diachronie longue, tant sur le plan linguistique (lexique, syntaxe, morphologie, graphie, pragmatique, etc.) que sur le plan stylistique (repérage des évolutions des stylèmes et des phraséologismes) ou littéraire (repérage des topiques ou des motifs narratifs) ?

- Quels sont les modes d'interrogation choisis parmi les multiples possibilités offertes par l'outil adopté ?

- Quels méthodes et outils spécifiques ont été développés en vue de l'exploitation de corpus en diachronie longue ? les propositions pourraient s'axer par exemple sur les techniques de périodisation automatique (Gries & Hilpert, 2008), sur des indicateurs textométriques permettant de mesurer des tendances (Herman & Kovář, 2013 ; Hilpert & Gries, 2009 : 388-390), sur des caractéristiques chronologiques spécifiques (Salem, 2021 ; Lebart et *alii*, 1998 : 155-161 ; Diwersy et *alii*, 2021), ou sur de nouvelles méthodes textométriques dédiées à l'étude diachronique. On pourra également détailler des outils d'exploration et de visualisation originaux.

BUILDING LONG-DIACHRONY CORPORA: METHODOLOGIES, GOALS AND LINGUISTIC OR STYLISTIC RESEARCH

Julie Sorba, Univ. Grenoble Alpes, LiDiLEM

Olivier Kraif, Univ. Grenoble Alpes, LiDiLEM

Adam Renwick, Univ. Grenoble Alpes, LiDiLEM

Corinne Denoyelle, Univ. Grenoble Alpes, Litt&Arts UMR 5316

In recent decades, the digitisation of printed works and progress in NLP have significantly changed the way that corpora can be created and the way that research on these corpora has been carried out. It is now possible to obtain vast amounts of quantitative data which allow for fine-grained analysis and identification of linguistic or stylistic phenomena in written corpora of historical states of language. Digital corpora created over the last quarter century allow for an easier appreciation of the dynamics of French in the long-term: the *Grande Grammaire Historique du Français* (Marchello-Nizia *et alii*, 2020), completed after many years of work is

a shining example. We define long-diachronic corpora as periodised corpora, containing texts chosen for their representativeness of certain states of language (from Old French to Contemporary French, for example) corresponding to the time-periods covered by the corpus.

Since the 1980s, researchers have been able to benefit from *Frantext*, the first corpus of French language texts, which allowed for research on literary texts over a very large timespan. The pioneering work of the *Base de Français Médiéval* (1989, BFM) led to the creation of a corpus of literary and non-literary texts, albeit limited, as its name suggests, to Old and Middle French. A new version of the BFM database will be available in late 2022. Numerous further corpora restricted to specific genres or certain states of language have since been created (for example, the *Condé* project's corpus of Norman *coutumiers* spanning six centuries, the Synctactic Reference Corpus of Medieval French (SRCMF) and the *Sermo* project's corpus of XVIth -XVIIIth century protestant sermons). This current dynamic in the constitution and exploitation of diachronic corpora was, for example, at the heart of the international conference *ConCorDiaL: Building Long-Diachrony Corpora* , organized at the University of Grenoble Alpes (October 2022), and will be addressed in April 2023 in the workshop *Background noise or added value? Managing noise, NLP and corpus linguistics* (University Grenoble Alpes/University of Roma La Sapienza) and in 2024 in the conference *Tracing the Curve of evolution : Syntactic change through text types* (University of Caen, March 2024).

As Reppen (2010: 31) and Nelson (2010: 53) underline, the first step in building a corpus is defining the goal the corpus serves. For example, selecting comparable sources to allow for homogeneous quantitative analyses is essential and the timespan examined depends on the phenomena to be investigated (*GGHF* 2020: 43). The building of a corpus is thus the product of a set of reasoned decisions seeking to satisfy the representativeness principle according to which a corpus is "a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis" (Tognini-Bonelli, 2001: 2). This representativeness principle also takes account of the variety of different purposes for which corpora can be constructed: representativeness requirements will differ between lexicographers seeking to take account of the meaning of lexical units on the one hand, and stylists characterising a textual genre on the other. For some, it is essential that analyses should only be made of texts in their entirety (Rastier, 2011:33) whereas for others, a corpus can only ever be a sample of the phenomena analysed and can thus be built on samples of texts (Renouf, 1987, Biber, 1993). The goal of this issue of *Corpus* is to question not only the choice of what sources a long-diachronic corpus should include, but also the linguistic, stylistic or literary objectives that determine the contents of the corpus.

The proposed themes of this issue of *Corpus* take viewpoints that are retrospective (what have diachronic corpora shown? How can corpora built in recent decades be put to better use?) as well as forward-looking (what theoretical and methodological challenges await research based on diachronic corpora in the era of digital humanities and corpus-tool platforms?). Contributions can be based on both French-language and foreign-language corpora.

Theme 1: Building a corpus

Creating corpora capable of providing data on long timescales poses new questions of homogeneity of tools and formats at all stages of the preparation of the corpus, from the selection of texts to the precise manner in with they are processed. For example, in the presentation of the criteria used to construct the corpus for the *GGHF*, Prévost (2020 : 42-43), distinguishes between texts selected according to *paratextual* criteria "which have more to do with the modern speaker's point of view on the texts" and which involve the choice of reference texts such as la *Chanson de Roland* or the *Queste del Saint Graal*, and texts selected based on *descriptive* criteria which have more to do with the period specific to each linguistic phenomenon. In particular, contributions may focus on:

- the diversity or homogeneity of texts, at different hierarchical levels (domains, discourses, genres; on these categories, see *inter alios*, Malrieu & Rastier, 2001; Marchello-Nizia et al., 2020) or different types of variation (diatopic, diastratic)
- the origin of texts to be included in the corpus, depending on whether the corpus itself is based on secondary sources (previously published texts) or primary sources (yet-unpublished texts). When corpora are based on previously published texts, what can be done to compensate for the inevitable array of different editorial decisions? In the case of primary sources, what changes should be made at the written level, given the evolution of philological editorial practices over the centuries with regards to matters such as the segmentation of words, spelling, accents, punctuation, capitalisation?)
- types of coding and annotation to be added to the chosen texts (what types of additional information have been and should be preferred when enriching texts? How many layers of annotation were added?)

Theme 2: Doing research with after building corpora

As the goal of a corpus influences its composition and construction, questions should be raised about the data that is to be extracted from the corpus.

- What type of research do long-diachronic corpora allow, be it on the linguistic level (lexicon, syntax, morphology, orthography, pragmatics…), the stylistic level (identifying changes in stylistic features and phraseological units…) or the literary level (identification of narrative topics or motifs)?
- Which ways of consulting the corpus were chosen from among the multiple possibilities offered by the chosen tools?
- What methods and specific tools have been developed to facilitate the analysis of long-diachronic corpora? Proposals can address techniques automating the division of corpora into stages (Gries & Hilpert, 2008), trend detection and measurement (Herman & Kovář, 2013; Hilpert & Gries, 2009: 388-390), specific chronological characteristics (Salem, 2021 ; Lebart et al. 1998: 155-161; Diwersy et al., 2021), as well as new textometric methods for the study of diachronic corpora. Proposals may additionally analyse novel corpus exploration and visualisation tools.

CALENDRIER / CALENDAR

Lancement de l'appel : 16 novembre 2022
Launch of the CFP : 16 November 2022

La soumission se fait en trois étapes (langues acceptées : français et anglais)
The submission process of articles, written in either French or English, consists of three stages

1. Résumés (500 mots, bibliographie non comprise), date limite : **15 janvier 2023**
Abstracts: 500 words, excluding references, deadline: **15 January 2023**
Notification aux auteurs (résumés) : 30 janvier 2023
Confirmation of acceptance (abstracts): 30 January 2023

2. Article complet (env. 30 000 signes espaces compris), date limite : **16 avril 2023**
Full paper submission (article length: 30,000 signs including spaces), deadline: **16 April 2023**
Notification aux auteurs (articles expertisés) : 30 juin 2023
Confirmation of acceptance (peer-reviewed papers): 30 June 2023

3. Article révisé (mise en page minimaliste, corps Times New Roman 12, titres en gras, format .doc(x) ou .dot), date limite : **1er octobre 2023**
Revised article (.doc(x) or .dot file format, minimal page formating (Times New Roman 12, bold headings), deadline: **1st October 2023**

Publication : janvier 2024
Publication: January 2024

Contact : julie.sorba@univ-grenoble-alpes.fr ; corinne.denoyelle@univ-grenoble-alpes.fr ; olivier.kraif@univ-grenoble-alpes.fr ; adam.renwick@univ-grenoble-alpes.fr

REFERENCES

BIBER D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4): 243-257.
DIWERSY S., JACKIEWICZ A., LUXARDO G. & STEUCKARDT A. (2021). Les sens de « numérique » : émergence d'emplois et dynamique du changement sémantique. *Linx*82. https://doi.org/10.4000/linx.8153
GALLERON I., FATIHA I., LAVRENTIEV A., DEMONET M.-L. & RÉACH-NGÔ A. (2021). Décrire les textes dans le cadre d'une édition numérique : Le thésaurus "Typologie textuelle" du Consortium CAHIER.
GLIKMAN J. & VERJANS T. (dir.) (2021). Regards linguistiques sur les éditions de textes médiévaux, *Diachroniques*, 8 : 7-16.
GRIES S. Th. & HILPERT M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3: 59–81.
HERMAN O. & KOVÁŘ V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing*, *RASLAN* 2013: 79–85.

HILPERT, M. & GRIES, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4): 385–401.
LAVRENTIEV A., GUILLOT-BARBANCE C. & HEIDEN S. (2021). Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots, *Diachroniques*, 8 : 76-102.
LEBART L., SALEM A. & BERRY L. (1998). *Exploring Textual Data*. Kluwer Academic Publisher.
MALRIEU D. & RASTIER F. (2001). Genres et variations morphosyntaxiques. *Traitement automatique des langues*, 42.2 : 547-577.
MARCHELLO-NIZIA C., COMBETTES B., SCHEER T. & PRÉVOST S (2020). *Grande Grammaire Historique du Français (GGHF)*. De Gruyter.
MARTINEAU F. (2008). Un corpus pour l'analyse de la variation et du changement linguistique, *Corpus*, 7 <https://doi.org/10.4000/corpus.1508>
MARTINEAU F. & SEGUIN M.-C. (2016). Le Corpus FRAN : réseaux et maillages en Amérique française, *Corpus*, 15 <https://doi.org/10.4000/corpus.2925>
MCENERY T. & WILSON A. (dir.) (2001). *Corpus linguistics*, Edinburgh University Press.
NELSON M. (2010). Building a written corpus. In A. O'Keeffe & M. Mc Carthy (éd.), *The Routledge Handbook of Corpus Linguistics* (p.53-65). Routledge.
PREVOST S. (2015). Diachronie du français et linguistique de corpus : une approche quantitative renouvelée. *Langages*, 197 : 23-45 <https://doi.org/10.3917/lang.197.0023>
RASTIER F. (2011). *La mesure et le grain. Sémantique de corpus*. Honoré Champion.
REPPEN R. (2010). Building a corpus. What are the key considerations? In A. O'Keeffe & M. Mc Carthy (éd.), *The Routledge Handbook of Corpus Linguistics* (p.31-37). Routledge.
SALEM A. (2021). Le temps lexical. *Histoire & Mesure*. Vol. XXXVI-2
TOGNINI-BONELLI E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company.
ZUFFEREY S. (2020). *Introduction à la linguistique de corpus*, ISTE Editions.